Milyen a jó Humor?

Novák Attila

MorphoLogic Kft., Budapest
novak@morphologic.hu

Kivonat. Magyar nyelvű szövegek morfológiai elemzésére elterjedten alkalmazzák a MorphoLogic Kft. által kifejlesztett Humor programot. Bár maga a program hatékony eszköznek bizonyult, a Humor adatbázisának formátumával problémák voltak a karbantarthatóság, az olvashatóság, a javíthatóság és a bővíthetőség szempontjából. Ez az előadás azt mutatja be, hogyan sikerült ezt a problémát az elemzőprogram módosítása nélkül a nyelvi adatbázis többszintűvé tételével orvosolni.
Kulcsszavak: automatikus morfológiai elemzés, nyelvi adatbázis.

A Humor morfológiai elemző

A magyarhoz hasonlóan bonyolult morfológiájú nyelvek számítógépes feldolgozása elképzelhetetlen hatékony morfológiai elemzőprogram nélkül. Magyar nyelvű szövegek morfológiai elemzésére Magyarországon leginkább a MorphoLogic Kft. által kifejlesztett Humor programot alkalmazzák (Prószéky és Kis, 1999). Ennek különböző változatait már több mint egy évtizede használják, és időközben a magyar mellett más nyelvekhez is készültek Humor alapú morfológiai elemzők. Bár maga a program hatékony eszköznek bizonyult, az elemző használhatóságát elsősorban az általa használt morfológiai adatbázis minősége határozza meg. Ez az előadás az elemző rövid ismertetése után egy olyan nyelviadatbázis-leíró rendszert mutat be, melynek segítségével jó minőségű magyar morfológiai adatbázist hoztunk létre a Humor elemzőhöz.

A Humor elemző jellemzői

A program klasszikus 'item-and-arrangement' típusú elemzést hajt végre (Hockett, 1954): egy szóalak lehetséges elemzéseit morfsorozatokként adja meg. A szót felépítő minden morfnak kiírja a felszíni és mögöttes alakját, valamint a kategóriáját (amely strukturált információt is tartalmazhat, de lehet belső szerkezet nélküli címke is). Az utóbbi kettő alapján általában azonosítható, hogy melyik morfémáról van szó. Azoknak a homonim lexémáknak az esetében, ahol a szófaj megadása nem elegendő az egyértelműsítéshez, azt a megoldást választottuk, hogy a lexikai alakhoz egyértelműsítő indexet toldottunk (pl. *szél_légmozgás*/*szél_perem*).

A program belső összetevős szerkezet nélküli lapos morfsorozatokként elemzi a szavakat. Ennek az az oka, hogy a program reguláris szónyelvtant tartalmaz, amely determinisztikus és epszilonmentes véges állapotú automataként van implementálva.

Ez egyrészt jóval gyorsabb, mint egy környezetfüggő nyelvtanon alapuló elemző, másrészt ezzel a megoldással elkerüljük sok irreleváns szerkezeti többértelműség előállítását, amit a megfelelő környezetfüggő elemző generálna (pl. a többszörösen képzett összetett szavak esetében).

Az elemző működése

Az elemző mélységi keresést végez a beadott szóalakon a lehetséges elemzések után. Olyan morfokat keres a szótárában, amelyeknek a felszíni alakja illeszkedik a megadott szó még elemzetlen részére. A lexikon nemcsak morfokat, hanem morfsorozatokat is tartalmazhat, amelyeket az elemző így egy lépésben ismer fel.

Elemzés közben a program kétféle ellenőrzést hajt végre. Egyrészt lokális kompatibilitás-ellenőrzést végez az egymás mellett álló morfok között: ellenőrzi a morfofonológiai és a lokálisan ellenőrizhető morfotaktikai feltételek teljesülését. Az előbbire példa a magyarban a magánhangzó-harmónia, az utóbbira pedig az a megszorítás, hogy névszói toldalékok csak névszótöveket követhetnek. Másrészt azt is ellenőrzi, hogy az elemzést alkotó morfémák a nyelv lehetséges szókonstrukciói egyikét testesítik-e meg (megfelelnek-e az adott nyelv morfológiai konstrukcióit leíró szónyelvtannak). A magyarban például a tő+képzők+ragok alakú morfémasorozatok jól formáltak, ugyanilyen kategóriájú morfémák más sorrendben azonban nem jók. A szónyelvtan nem szomszédos összetevők közötti megszorítások ellenőrzését is lehetővé teszi: pl. a *leg-* felsőfokjelet egy tőle jobbra álló morfémának (leggyakrabban a *-bb* középfokjelnek) engedélyeznie kell, közöttük azonban számos más morféma is állhat.

A Humor nyelvi adatbázisa

A program hatékony működésének az a feltétele, hogy az elemzés közben végrehajtandó ellenőrzések nagyon egyszerű és gyors műveletek legyenek. Ehhez az kell, hogy az adatbázis rengeteg redundáns információt tartalmazzon explicit formában, hogy ezeket ne elemzés közben kelljen kiszámítani. A legfőbb probléma az volt, hogy a MorphoLogicnak nem voltak eszközei az elemző által használt adatbázist alkotó redundáns adatszerkezetek létrehozására és karbantartására. A szomszédos morfok közötti lokális kompatibilitás-ellenőrzéshez használt adatszerkezeteket, az allomorfok (és nem morfémák) leírását tartalmazó lexikonokat és a szónyelvtant definiáló véges állapotú automata leírását egyszerű szövegszerkesztő segítségével kellett létrehozni és karbantartani.

A gép számára optimalizált leírások az emberek számára lényegében olvashatatlanok, és ezért nagyon nehéz őket konzisztens módon karbantartani, módosítani, a hibákat megtalálni és kijavítani. A Humor például kétféle adatszerkezetet használ a lokális kompatibilitás ellenőrzésére: egyrészt bináris tulajdonságvektorokat, másrészt kompatibilitási mátrixokat. Mindkét adatszerkezet nagyon nehezen olvasható és a mátrixok kézzel való konzisztens módosítása lényegében lehetetlennek bizonyult. Ráadásul ha egy tulajdonságot vagy jelenséget (pl. a magánhangzó-harmóniát) egyszer az egyik adatszerkezettel ábrázoltunk, nagyon nehéz áttérni a másik adatszerkezettel

való ábrázolásra. Ennek az volt a következménye, hogy a leírások a fejlesztők legjobb szándéka ellenére is hibásak és inkonzisztensek maradtak.

Ezt a problémát az elemzőprogram módosítása nélkül, a nyelvi adatbázis többszintűvé tételével sikerült orvosolni. Egy olyan nyelviadatbázis-leíró keretrendszert hoztunk létre, amelyben a nyelvész magas szintű, ember számára olvasható formátumú leírást készíthet a leírandó nyelv morfológiájáról. Ez a leírás morfémák és nem allomorfok leírását tartalmazza, és az egyes morfémáknak csak azok a tulajdonságai szerepelnek benne, amelyek nem megjósolhatóak. Mivel ez a reprezentáció nem tartalmaz redundáns információt, könnyű konzisztens állapotban tartani. A leírásnak ezen a magas szintjén könnyen lehet a lexikont bővíteni és javítani. Ebből a leírásból a nyelvész által definiált szabályok alapján a keretrendszer állítja elő azokat a redundáns adatszerkezeteket, amelyeket az elemző használ.

A szóalaktani adatbázis létrehozása

A nyelviadatbázis-leíró keretrendszert használó nyelvész munkája a következő feladatok elvégzéséből áll:

- A nyelv morfémakategória-készletének leírása (szófajok, toldalékkategóriák).
- A tő- és toldalékalternációk megadása: le kell írni azt a műveletet, amellyel az egyes fonológiai allomorfiaosztályokba tartozó tövek lexikai alakjából az egyes allomorfok előállnak. Ennek leírására a keretrendszerben reguláris kifejezéseket lehet használni. Meg kell állapítani, hogy mely morfok váltják ki a váltakozást. Ha a váltakozásnak fonológiai vagy fonotaktikai feltétele van, akkor közvetlenül ezekre a tulajdonságokra lehet hivatkozni. Ha idioszinkratikus lexikai jegyek is szerepet játszanak, akkor ezeket be kell vezetni.
- A morfológiai tulajdonságok feltérképezése: azonosítani kell minden olyan tulajdonságot, amely a nyelv morfológiájának leírásánál szerepet játszik. Ezek különbözőfélék lehetnek: vonatkozhatnak a morféma kategóriájára, egy allomorf hangalakjának, illetve írott alakjának valamilyen morfológiailag releváns jellemzőjére, vagy a morféma által kiváltott idioszinkratikus váltakozásra (pl. tőalternációkra).
- A szomszédos morfok közötti szelekciós megszorítások definiálása: ezeket a megszorításokat egy olyan követelményformula formájában kell leírni, amelyet bármely, a morffal szomszédos más morf tulajdonsághalmazának ki kell elégítenie. A tulajdonsághalmazok és a követelményeket leíró formulák az előző pontban azonosított morfológiai tulajdonságokat tartalmazhatják. Minden morf két tulajdonsághalmazzal rendelkezik: az egyiket a morffal balról, a másikat a morffal jobbról szomszédos morfok látják. Hasonlóképpen minden morf egy-egy formulával megszorítást tehet mind a vele balról mind a vele jobbról szomszédos morfémákra nézve. Egy morfot csak akkor követhet egy másik, ha mind a bal oldali morf jobbról látható tulajdonságegyüttese kielégíti a jobb oldalinak a bal szomszédjával szemben támasztott követelményeit, mind pedig a jobb morf balról látható tulajdonságegyüttese kielégíti a bal oldalinak a jobb szomszédjával szemben támasztott követelményeit.
- A morfémák és allomorfok tulajdonságai közötti implikációs viszonyok megadása: ezeket az implikációs viszonyokat olyan szabályok formájában kell megfogalmazni, amelyek leírják, hogy az allomorfok redundáns tulajdonságai hogyan számít-

hatók ki a már ismert (a lexikonban megadott, vagy korábban már kiszámított) tulajdonságaikból (ide értve az alakjukat is). A szabályok default tulajdonságokat is bevezethetnek mind a morfémák mind az allomorfok szintjén, és a szomszédos morfokra vonatkozó megszorításokat is megfogalmazhatnak. A szabályokat egy erre a célra alkotott viszonylag egyszerű procedurális nyelven lehet leírni. A tő- és toldalékallomorfok előállítását leíró mintákat is a szabályfájlok tartalmazzák.

- A tő- és toldaléklexikonok előállítása: a morfológiai elemző által használt lexikonnal ellentétben a nyelvész által létrehozott lexikonok morfémák és nem allomorfok leírását tartalmazzák. A morfémákat a lexikai alakjuk, a kategóriájuk és a megjósolhatatlan vagy rendhagyó tulajdonságaik és elvárásaik megadásával kell leírni. A rendhagyó toldalékolt alakok és szuppletív allomorfok is megadhatók a lexikonban. Ezek leírásának ez a preferált módja, bár a rendszer azt is lehetővé teszi, hogy nagyon szűk körben működő szabályokkal állítsuk őket elő. A komplex lexikai egységek (elsősorban az összetett szavak) konzisztens és gazdaságos leírásának elősegítésére beépítettünk a rendszerbe egy egyszerű öröklési mechanizmust, amelynek segítségével az összetett lexikai egységek alapesetben az utótagjuktól öröklik a tulajdonságaikat. Az öröklési mechanizmus működésének az a feltétele, hogy a szavakat az összetételi határok jelölésével kell a lexikonba felvenni.
- A szónyelvtan leírása: a szavak belső alaktani szerkezetére vonatkozó megszorításokat (ideértve a nem szomszédos morfémák közötti megszorításokat is) a szónyelvtan írja le. A Humor elemző reguláris szónyelvtan használatát teszi lehetővé. A nyelvtant az elemző számára determinisztikus, epszilonmentes kiterjesztett véges állapotú automata formájában kell leírni. Az automata annyiban kiterjesztett, hogy az egyes állapotátmenetek megadásakor egy véges bináris vagy több bites változókészlet elemeinek értékét lehet módosítani, illetve ellenőrizni. A keretrendszer az automata leírását egyrészt azzal könnyíti meg, hogy szimbolikus változónevek definiálását teszi lehetővé, és ezzel olvashatóbbá teszi a leírást, másrészt egy hatékonyan használható makródefiniáló és –kezelő eszközt is biztosít, amelynek segítségével több hasonló, de részleteiben különböző állapotátmenetet lehet egyszerre definiálni (ami a bonyolultabb automaták leírását nagyban megkönnyíti).
- Külön toldaléknyelvtan létrehozása (nem kötelező): egy irányított gráf formájában külön toldaléknyelvtant lehet definiálni, amelynek felhasználásával a keretrendszer a toldaléklexikonból elemzett toldaléksorozatokat állít elő. Ezeknek az előre megelemzett morfsorozatoknak az elemző lexikonjába való felvétele jelentősen gyorsítja az elemző működését, mert a magyarban és a hozzá hasonló agglutináló nyelvekben nem ritkák a hosszú toldaléksorozatok. A toldaléknyelvtan használatának a másik előnye az, hogy a szónyelvtannak azt a részét, amit a toldaléknyelvtan segítségével leírtunk általában ki lehet hagyni az elemző által használt szónyelvtan-leírásból, aminek eredményeképpen az utóbbi jelentősen egyszerűsödik.

A morfológiai adatbázis átalakítása

A fent leírt módon elkészített leírás alapján a keretrendszer olyan reprezentációt hoz létre, amelyben már minden morféma minden allomorfja az összes tulajdonságával és elvárásával együtt explicit módon szerepel. Az így előálló reprezentáció még mindig olvasható formában tartalmazza az egyes morfok tulajdonságait és szelekciós megszo-

rításait kifejező formulákat, így a nyelvész könnyen ellenőrizheti a leírások helyességét. Az alábbi példa a *kutya* szó redundáns reprezentációját mutatja be.

```
lemma:  'kutya[FN]'
 root:  'kutya'
 allomf:'kutya'
  mcat: 'S_FN'
  rp:   '-Vs -nyi -sÁg -tAlAn =_s =_t =i =jA =vAl VHB
Vfin cat_N cmp2 sfxable mcat_stem'
  rr:   '!FVL'
  lp:   'Cini comp2 k_ini'
  lr:   '!cat_vrb'
 allomf:'kutyá'
  mcat: 'S_FN'
  rp:   '-Vs -nyi -sÁg -tAlAn =_s =_t =i =jA =vAl VHB
Vfin cat_N cmp2 sfxable mcat_stem'
  rr:   'FVL'
  lp:   'Cini comp2 k_ini'
  lr:   '!cat_vrb'
```

A *kutya* tőnek, amely főnév (`[FN]`) kategóriájú két alakja (allomorfja) van: egy *kutya* és egy *kutyá* alakú. A két allomorf jobb, és bal oldali tulajdonságai (`rp` ='right side properties', ill. `lp`='left side properties') valamint a bal oldali elvárásaik (`lr`='left side requirements') is megegyeznek. A jobb oldali tulajdonságok közül a – kezdetűek arra utalnak, hogy a megfelelő képzőt a tő felveheti, az = kezdetű tulajdonságok azt írják le, hogy a megfelelő toldalékot a tő milyen alakban veszi fel. A `Vfin`, `Cini`, `k_ini` a morf alaki tulajdonságait írják le (magánhangzóra végződik, mássalhangzó kezdetű, *k* kezdetű), a `VHB` azt írja le, hogy a harmonikus toldalékok hátul képzett változata kapcsolható hozzá, a `cat_N`, `cmp2`, `sfxable`, `mcat_stem` pedig a morféma kategoriális tulajdonságait írják le (főnév, szerepelhet összetétel második tagjaként, toldalékolható és tő), amelyek – az elemző számára készített redundáns leírásról lévén szó – minden allomorf leírásánál explicit módon megjelennek. A ! a tagadás jele: a `!cat_vrb` megszorítás jelentése: igető után nem állhat. A *kutyá* allomorf jobb oldali szomszédainak `FVL` ('final vowel lengthening') tulajdonsággal kell rendelkezniük, vagyis olyan toldaléknak kell lenniük, amelyik kiváltja a tővégi alsó magánhangzó (*a* vagy *e*) megnyúlását. A *kutya* allomorftól jobbra éppen az ilyen tulajdonsággal bíró morfok nem állhatnak (`!FVL` megszorítás).

Ezt a reprezentációt a keretrendszer a következő lépésben az elemző által használt formájúra alakítja. A fordítás alapjául egy olyan leírás szolgál, amely minden egyes, a nyelv leírásánál használt tulajdonságra megadja a kódolás módját az elemző számára. Lehetőség van arra is, hogy egy tulajdonságot a fordításkor figyelmen kívül hagyjunk, így létre lehet hozni az elemző olyan módosított változatait is, amelyek bizonyos megszorításokat figyelmen kívül hagynak, és ily módon túlelemeznek. A fordítás alapjául szolgáló leírás elkészítése szintén a keretrendszer felhasználójának a feladata.

Az általunk használt egyszerű propozicionális leírás minden tulajdonságot binárisan reprezentál, a leírandó nyelv morfológiája azonban olyan, hogy bizonyos tulajdonságok igaz voltából automatikusan következik, hogy egyes más tulajdonságok nem lehetnek igazak az adott objektumra, ha pl. egy tő ige, akkor nem lehet főnév is egyben. A keretrendszer lehetővé teszi, hogy kifejezzük, hogy bizonyos tulajdonsá-

gok ugyanannak a jegynek egymást kizáró lehetséges értékei. Az ilyen tulajdonságokat valódi független bináris tulajdonságokra dekomponálhatjuk, ami egy konjunktív következményformula (tkp. egy jelentésposztulátum) formájában adható meg a tulajdonság kódolását megadó leírásban.

Az új magyar morfológiai adatbázis

A keretrendszer felhasználásával teljesen új leírást készítettünk a magyar morfológiáról. Az eredeti Humor adatbázisból kizárólag lexikai információt vettünk át: az új elemző tőmorféma-készlete eleinte megegyezett az eredetiével, de rengeteg hibás vagy inkonzisztens kategóriacímkét kijavítottunk, és a komplex (összetett, képzett) tövek szegmentálását megadtuk (erre az öröklési mechanizmus működéséhez is szükség van). A zárt tőosztályokba tartozásra vonatkozó információt (pl. *v*-vel bővülés, tőmagánhangzó-rövidülés, nyitótőség stb.) szintén az eredeti adatbázisból nyertük (javításokkal).

A toldalékok kategóriacímkéi – a kompatibilitás kedvéért – általában megegyeznek a korábbiakkal, de néhány korábban szételemzett toldalékot atominak tekintettünk az új leírásban (pl. a *–hAtÓ* és a *-hAtAtlAn*). A névmás, mint kategória megszűnt: a névszói és határozói kategóriákon belül vannak névmási tulajdonsággal bíró tövek.

Paradigmatikus információt egyáltalán nem vettünk át az eredeti leírásból; a paradigmák az allomorfokat és tulajdonságaikat, illetve elvárásaikat kiszámító szabályrendszer révén állnak elő.

Az eredeti rendszerrel ellentétben az újba nagyon könnyű új szavakat felvenni, mert csak azokat a megjósolhatatlan tulajdonságaikat kell a szótárba felvenni, amelyek különböznek a defaulttól. Ez a szavak túlnyomó többsége esetében a lexikai alakra és a kategóriacímkére korlátozódik, illetve az esetleges összetételi határok megadására (a *kutya* szó reprezentációja a tőadatbázisban például egyszerűen `kutya[FN]`, ebből automatikusan áll elő a fentebb látott redundáns reprezentáció).

A keretrendszer használatával készült egyébként egy jó minőségű spanyol morfológiai elemző is, ezen kívül egy folyamatban lévő projekt keretében számos kisebb finnugor és más uráli nyelv leírására is ezt a rendszert használjuk.

Hivatkozások

C. Hockett. 1954. Two models of grammatical description. *Word* 10 (2): 210–234.

Prószéky Gábor és Kis Balázs. 1999. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 261–268. College Park, Maryland, USA